基于深度学习的广告CTR预估算法

摘要：本文主要介绍了广告CTR预估算法在引入深度学习之后的基本演化过程及一些最新的进展，重点是从工业实现和应用的视角对Deep CTR模型进行剖析，探讨为什么这样设计模型、模型的关键要点是什么。主要内容按照“内”、“外”两个不同的角度进行介绍：内部集中介绍了典型模型的网络结构演化过程，外部则关注于不同数据、场景和功能模块下模型的设计思路。

演讲嘉宾简介：

朱小强，花名怀人，阿里妈妈高级算法专家，领导了核心的排序算法与机器学习平台团队，负责阿里精准展示广告的CTR/CVR预估系统/算法和架构的设计优化、大规模分布式机器学习/深度学习平台建设等工作。

[本次直播视频精彩回顾，戳这里！]

[本次直播视频PDF下载！]

以下内容根据演讲嘉宾视频分享以及PPT整理而成。

本次的分享主要围绕以下三个方面：

一、CTR预估问题的特点与挑战——以阿里定向广告为例 网络爬虫技术入门

二、基于深度学习的CTR预估算法演化——内外兼修之道

三、总结与展望——新的起点

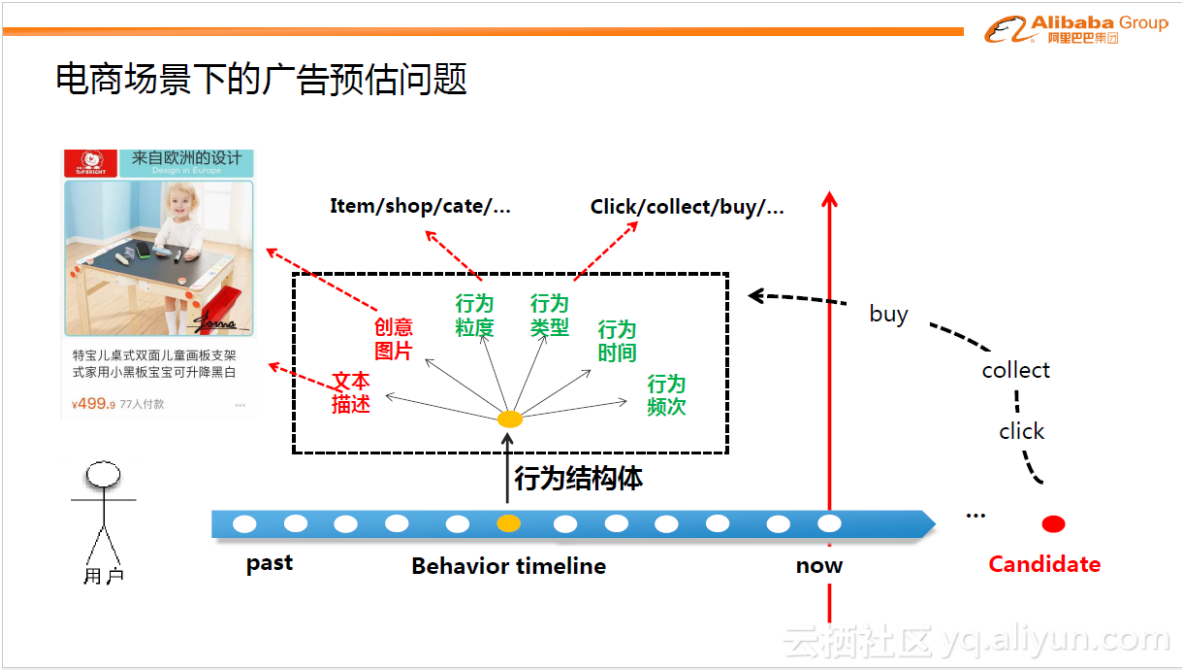
一、CTR预估问题的特点与挑战——以阿里定向广告为例

下图中可以看到手机淘宝端的定向广告形态。左边是首焦场景，在淘宝顶端的位置会有浮动的Banner广告。右边是往下滑动时候的导购场景（猜你喜欢区块），投放的是Item广告。这些不同形态的定向广告背后其实有一些内在的、从machine learning视角来看相似的特征。简单来说，可以归纳为几个方面，一个方面是广告中展现的创意图片，第二个是图片的文字信息，还有一些在背后看不到摸不到的统一的ID体系，比如某件商品是什么商品，属于哪个品牌等等信息。定向广告复杂多样的富媒介形态以及高维海量数据空间，给广告点击率预估问题带来了不小的挑战。

下图是电商环境下CTR预估问题的数学化模拟。假设一位用户登录手机淘宝，我们首先可以拿到用户的一些历史行为数据，这些行为数据构成了我们对用户兴趣的表达刻画。那么下一步需要预估给用户展现某个候选商品candidate，用户发生点击/购买的概率是多少。那么如何实现预估？我们需要利用历史行为数据建模出用户的兴趣偏好。

将用户的行为按照时间排列，可以构成一个行为时间轴。每个时刻点可以称为行为结构体，它包含了一系列表征行为的关键信息：比如此刻的行为类型，点击or购买某个商品；某个商品的文本描述信息；对应的创意图片；行为发生的时间，行为发生的频次；或者行为背后的粒度体系是什么，对应的是什么商品、什么店铺以及什么品牌等等。

这些大量的行为信息可以足够表达用户的兴趣偏好。时间轴左边是历史的静态信息，称为feature；右边就是待预测的用户的未来行为，如点击行为（点击概率）、购买行为（购买概率）等等。电商场景下的广告预估问题相比于大家熟知的静态预估模型有更大的挑战。



第一个挑战，在淘宝端每天有数亿的用户会登陆，并产生大量的行为。同时我们有海量的商品候选集，在淘宝中有大概10亿到20亿的商品，当然聚焦到广告商品，可能会有所减少，但依然达到了千万的数量级。如此，广告预估问题就变成了数亿用户与千万商品配对的点击概率预估问题，规模极其庞大。第二个挑战，每个用户行为特征背后，有大量的信号源，比如图像信号、文字信号、品牌偏好信号等等，这些信号如何去捕捉，如何进行统一建模？第三个挑战，在电商场景下用户的行为非常丰富，反映出用户的兴趣多样多变，寻找与建模用户点击某个广告商品背后的规律是高度非线性的问题。

在点击率预估问题上，传统的解法一般采用逻辑回归模型（Logistic Regression，LR）。但可惜的是，当数据本身有非常强的非线性pattern时，传统的线性LR模型受到了很大的挑战。在阿里广告定向发展初期，也尝试探索过线性模型的可能性，但由于线性模型本身过于简化，算法的发挥空间上受到了极大的压制。

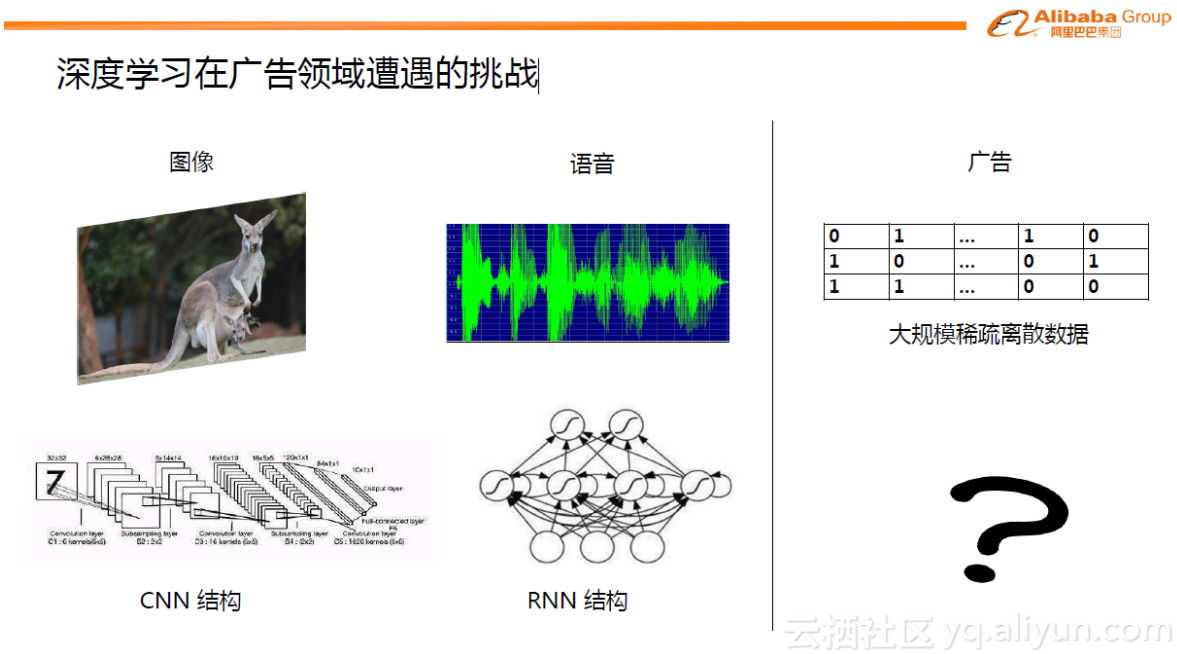
二、基于深度学习的CTR预估算法演化——内外兼修之道

2012年左右，当深度学习逐渐从学术界迁移到工业界时，阿里发现深度学习成为了一个有力的武器，可以更好地帮助广告CTR预估问题的求解。尤其是近几年里，基于Deep Learning的CTR预估算法正在经历快速的演变。下面主要介绍其演变背后的思路、核心关键以及具体可借鉴的要点。

1.深度学习在广告领域遭遇的挑战

传统的图像领域或NLP领域内，深度学习已经取得非常多的成果，在大多的问题上成为了state-of-the-art的方法。如图中显示，不同的领域有不同的适用结构。如图像领域内的CNN结构、语音的RNN结构。那么回到广告领域内，究竟什么结构是合适的？广告预估问题有很强的特点，它的特征极其的大规模和稀疏。典型的数量级从百万级，千万级到数亿级都有，而且大都是0或1这类没有直接意义的数据。

这个问题一度成为广告预估问题引入深入学习的关键点。

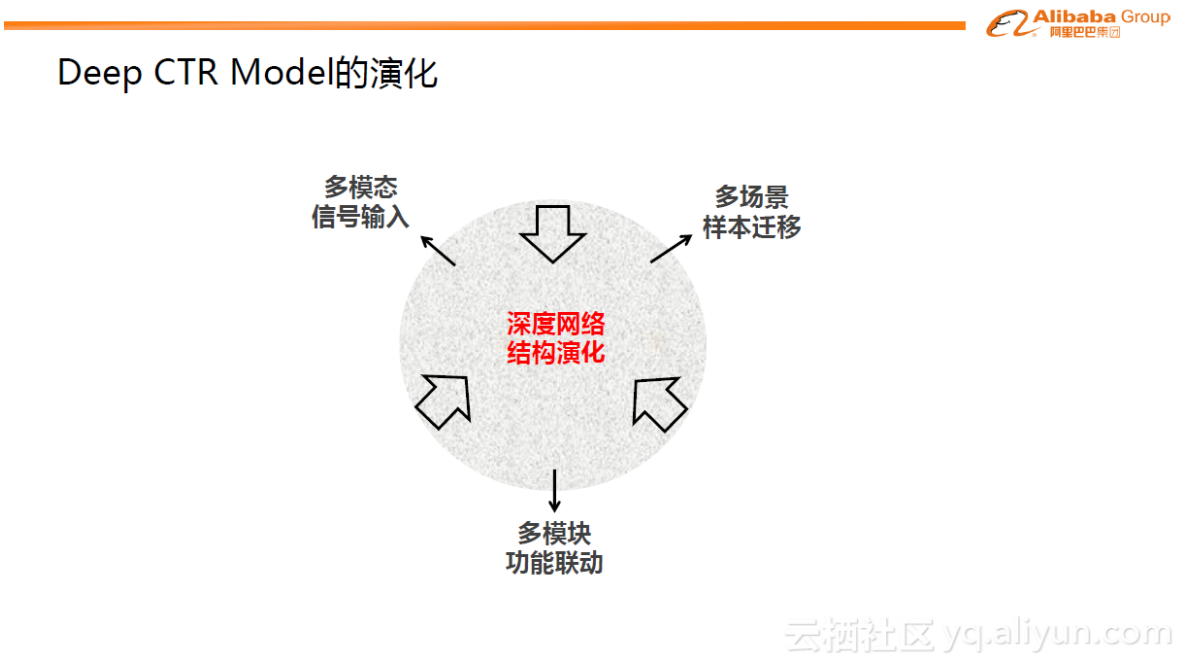


2.深度学习引入广告预估问题后的演化

下面从两个视角来介绍最近几年学术界和工业界对于上面问题的探讨，也叫做内外兼修。内部是指深度网络的核心结构如何演化，外部是指如何引入多模态的信号输入，多场景的样本迁移及多个模块的功能联动等等。

1）内部演化

下面主要介绍最近几年Deep CTR Model的内部结构上的演化，并进行简单梳理。

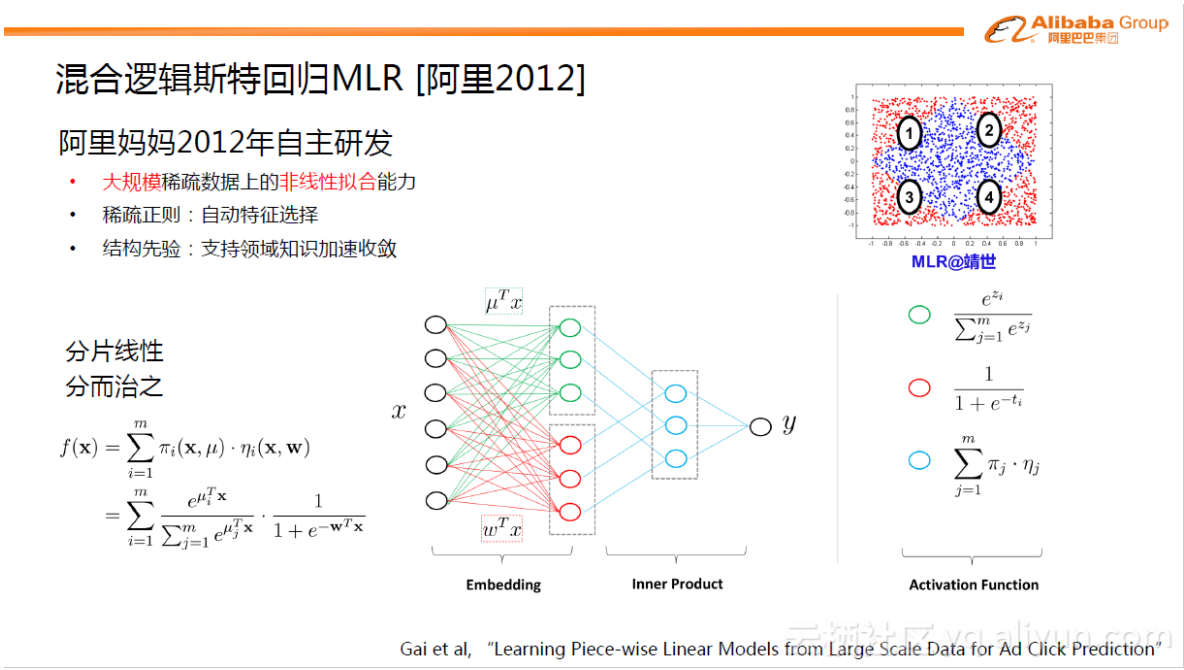


a. 混合Logistic Regression(MLR)模型

阿里自主研发的MLR模型是对线性LR模型的推广，它利用分片线性方式对数据进行拟合。基本思路是采用分而治之的策略：如果分类空间本身是非线性的，则按照合适的方式把空间分为多个区域，每个区域里面可以用线性的方式进行拟合，最后MLR的输出就变为了多个子区域预测值的加权平均。在今天看来，MLR模型是带有一个隐层的神经网络。

如下图，X是大规模的稀疏输入数据，MLR模型第一步是做了一个Embedding操作，分为两个部分，一种叫聚类Embedding（绿色），另一种是分类Embedding（红色）。两个投影都投到低维的空间，纬度为M，对应的是MLR模型中的分片数。完成投影之后，通过很简单的内积（Inner Product）操作便可以进行预测，得到输出Y。右边是不同节点上的激活函数Activation Function，已按不同颜色区分。

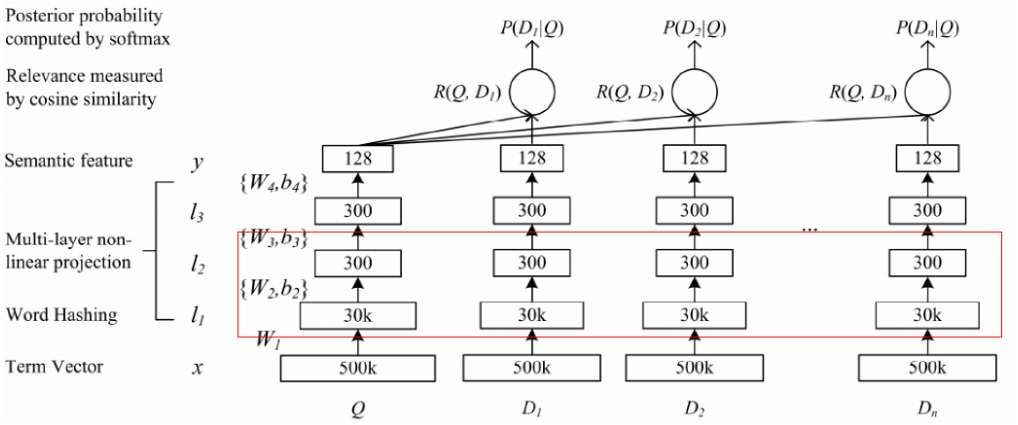
在2012年左右，MLR模型便在阿里的主流业务中进行服务，证明了其巨大的优越性。MLR模型最大的意义在于，它是首个在大规模稀疏数据上探索和实现了非线性拟合能力的模型，相关的细节内容可从论文中查询：Gai et al, “Learning Piece-wise Linear Models from Large Scale Data for Ad Click Prediction”



b.DSSM模型

如果说MLR模型是阿里巴巴初次对于深度学习方面的探索，在深度学习真正引入到广告预估问题中之后，出现了更多演变的模型。Deep Structured Semantic Model（DSSM）模型是微软2013年提出的。虽然在最初DSSM模型不是用于广告预估，但是现在看来，它为广告预估提供了一个很好的思路。这里主要关注下图中红色框内的部分，原理是把query/doc中的关键信息（Term Vector）提取出来进行简单的Word Hashing之后，把query/doc域分别投影到300维的子空间去。query里的每个word都对应一个300维的向量，一个query里会有多个向量，后面用sum求和操作得到一个汇总的300维向量，这是一个典型的Embedding操作。从图中可以看到，30k是指word字典的长度，300是embedding维数，30k*300≈千万量级的参数。DSSM模型第一次探索了如何把大量稀疏的ID进行稠密表达的路径。

DSSM模型 [微软2013]



Deep Structured Semantic Model(DSSM)

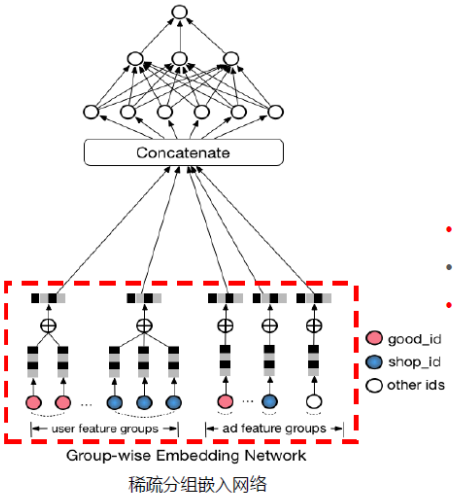
Huang et al, "Learning Deep Structured Semantic Models for Web Search using Clickthrough Data"

当然，DSSM模型本意不是用于广告预估问题。在深度学习最初引入CTR预估问题时，当时业界的一些公司如Google、百度等已经在探索如何把大量的稀疏数据进行降维的方法。一个典型的做法是用辅助的方式分两阶段进行：第一阶段，用FM模型把大量的稀疏ID学习到对应的embedding表达，跟DSSM模型类似，能够得到几百维的稠密向量。第二阶段是基于稠密的输入用多层全连接网络预测最后的目标。从今天的视角来看，这种两阶段的方式是不如整体的端到端模型的。这个思考点在2013年-2014年左右一直有人进行尝试，但当时一是因为深入学习框架的没能普及，二是对整个计算力的估计不足，因此没有达到比较好的进展，直到2016年左右，才有所突破，当然这里面很重要的一点是得益于优秀的深度学习框架如TensorFlow、MXNet等的开源和普及，进一步促进了整个工业界的思考和前进。

c.稀疏分组嵌入网络结构（GwEN）

下面以阿里2016年的网格框架为例进行介绍。整个稀疏分组嵌入网络结构（GwEN）分为两部分，如下图左边所示。第一部分，把大规模的稀疏特征ID用Embedding操作映射为低维稠密的Embedding向量，然后把每个特征组的 Embedding进行简单的sum或average的pooling操作，得到Group-wise的Embedding向量。第二部分，多个特征组的向量通过Concatenate操作连接在一起，构成原始样本的完整稠密表达，喂给后续的全连接层。

稀疏分组嵌入网络GwEN [阿里2016]



Group-wise Embedding Network

稀疏分组嵌入网络

0	1	...	1	0
1	0	...	0	1
1	1	...	0	0

高维稀疏离散输入~10^9

0.2	...	0.05
-0.8	...	0.3
1.2	...	-0.2

低维稠密连续表示~10^3

- 全空间表示学习 → 参数规模10^9x 10^3 = 10^{12} → 万亿量级
- 借鉴CNN网络的局部感知野(Local Receptive Field)思想
- 分组表示学习 → 10^2量级的分组 → 参数规模10^7x 10^1x 10^2 = 10^{10} → 百亿量级

Group-wise Embedding

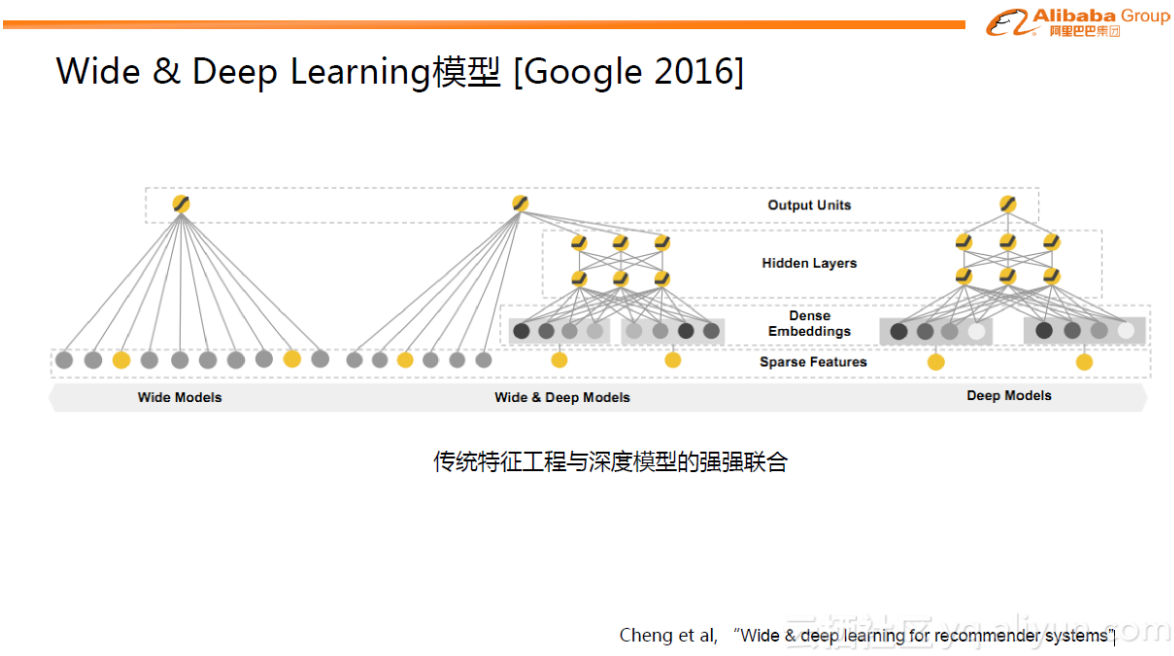
求解大规模稀疏数据问题的通用网络单元

Zhou et al, "Deep Interest Network for click-through rate prediction"

GwEN网络结构是比较基础的，但同样也非常重要。因为在最初大家普遍的直观思考是，假设有个高维的稀疏输入，典型的数量级达十的九次方，然后将每个ID学习到一个表达，如果表达太小，便不足以客刻画信息本身，那么设想投影维度控制在数百上千估计是合适的。早期百度或Google的探索中，大概是一样的量级。这样算来，全空间便达到万亿的量级，极其恐怖，一方面对于训练样本的要求，另一方面对于背后的计算能力的要求都非常高。Group-wise Embedding的核心想法是借鉴CNN网络的局部感知野（Local Receptive Field）的思想。以整体输出表达十的三次方向量为例，其实不需要每个ID达到千维的表达，因为在分组表示学习中每个特征组可以分别得到一个低维的表达，一共十的二次方量级分组，组里面的每个ID只需要十的一次方量级学习的表达即可。这样，整个参数规模可以直接压缩到百亿的量级，这是工业界比较舒服的量级。尽管GwEN这种网络结构非常简单，但提出了非常重要的Group-wise Embedding 的概念，现在也称为求解大规模稀疏数据问题的通用网络单元。GwEN网格结构在2016年左右在阿里内部已经上线。

d.Wide & Deep Learning模型

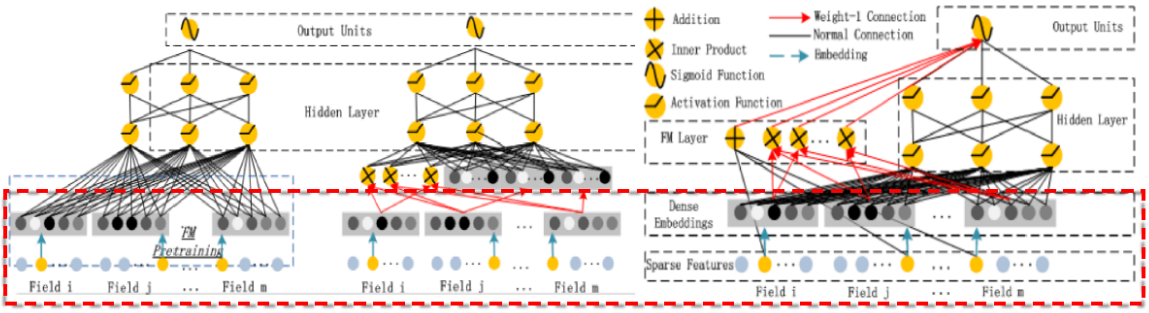
与阿里同时期，Google推出了Wide & Deep Learning（WDL）模型，一个非常出名的模型。详细内容可以从论文中查询Cheng et al, “Wide & deep learning for recommender systems” 。WDL模型也非常简单，但巧妙的将传统的特征工程与深度模型进行了强强联合。Wide部分是指人工先验的交叉特征，通过LR模型的形式做了直接的预测。右边是Deep部分，与GwEN网络结构一样，属于分组的学习方式。WDL相当于LR模型与GwEN结合训练的网络结构。



e.FNN/PNN/DeepFM模型

GwEN和WDL是目前比较常用的模型，非常简单，所有后续人们继续做了很多改进，例如FNN，PNN以及DeepFM等。这些模型基础的部分与上面的GwEN和WDL模型类似，即Group-wise Embedding。改进的地方主要在后面的部分，引入了代数式的先验pattern，如FM模式，比较简单直接，可以给MLP 提供先验的结构范式。虽然理论上说，MLP可以表达任意复杂的分类函数，但越泛化的表达，拟合到具体数据的特定模式越不容易，也就是著名的“No Free Lunch”定理。因此代数式的先验结构引入确实有助于帮助MLP更好的学习。当然从另外的视角看，这种设定的结构范式比较简单，过于底层，也使得学习本身比较低效。

FNN/PNN/DeepFM模型



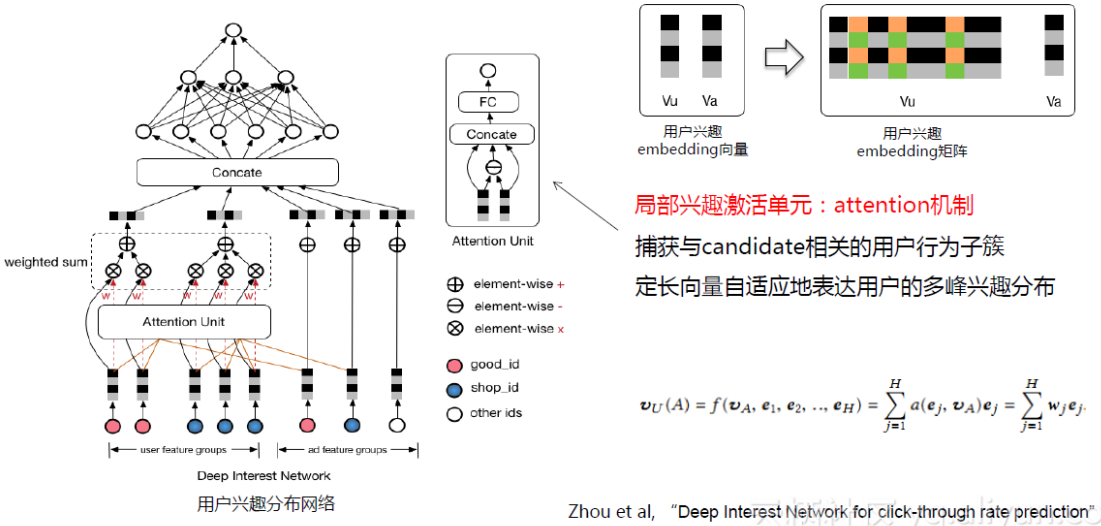
FNN [Zhang 2016] PNN [Qu 2016] DeepFM [Guo 2017]

Zhang et al, "Deep learning over multi-field categorical data - - A case study on user response prediction"
Qu et al, "Product based neural networks for user response prediction"
Guo et al, "DeepFM: A Factorization-Machine based Neural Network for CTR Prediction"

f.DIN模型

另外一个工作是阿里在2017年发表的用户兴趣分布网络DIN模型。与上面的FNN,PNN等引入低阶代数范式不同，DIN的核心是基于数据的内在特点，引入了更高阶的学习范式。互联网上用户兴趣是多种多样的，从数学的角度来看，用户的兴趣在兴趣空间是一个多峰分布。在预测多兴趣的用户点击某个商品的概率时，其实用户的很多兴趣跟候选商品是无关的，也就是说我们只需要考虑用户跟商品相关的局部兴趣。所以DIN网络结构引入了兴趣局部激活单元，它受attention机制启发，从用户大量的行为集合中捕获到与candidate商品相关的行为子簇，对于用户行为子簇，通过Embedding操作，做weighted sum便可很好的预估出用户与candidate相关的兴趣度。传统的GwEN、WDL、FNN等模型在刻画用户兴趣分布时，会简单的将用户兴趣特征组做sum或average的pooling操作，这会把用户真正相关的兴趣淹没在pooling过程中。DIN模型的具体细节可以参考论文：Zhou et al, "Deep Interest Network for click-through rate prediction" 。

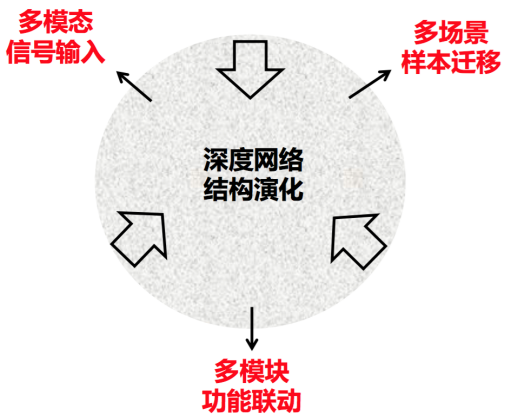
用户兴趣分布网络DIN [阿里2017]



Zhou et al, "Deep Interest Network for click-through rate prediction"

2)外部演化

Deep CTR Model的演化



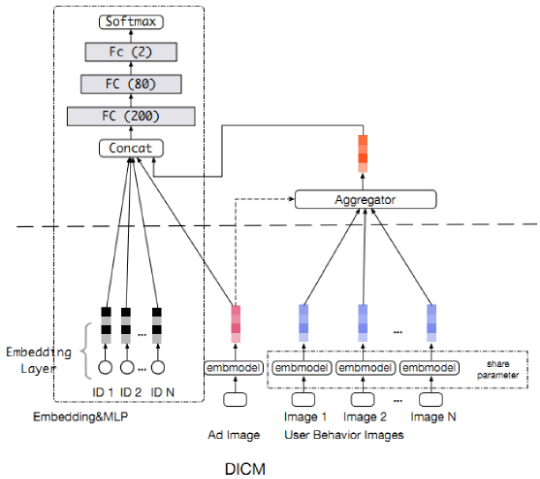
假设固定网络结构，那么外部演化则考虑的是有没有一些更好的特征输入或样本的方式可以帮助模型学习的更好。

a.多模态信号输入

首先介绍阿里在2017年发表的用户行为ID+图像的多模态学习模型的工作。用户在看到某个商品之后，映入眼帘的除了商品是什么的信息外，还有一系列的图片创意内容。比如下图是一款儿童画板商品，画板的大小、颜色甚至小宝宝可爱的模样等信息可能激发了用户点击的欲望。这类信息是无法单纯地通过画板这个ID完全表达的。所以这里面我们主要做的事情就是对于用户行为，除了商品ID之外，把对应的图像也放进来，统一表征用户的行为。

下图中左边部分描述了整个结构。与上面的DIN网络结构一致的是也使用了attention机制，引入Ad与User之间的相关性，不同的是网络结构将用户行为的ID特征与图片特征两种不同模态很自然的揉合在一起，解决了预测问题。这种做法在算法中非常直观，但事实上在背后真正建立模型时工程上面有很大的挑战。假设某个业务场景中有100亿的样本，每个样本有500个对应的用户行为ID特征，每个ID背后都有对应的图片。从图片视角来看，图片训练集有5亿张，多达8T数据，如果将图片训练装配到样本中平铺开来会接近800TB的数据。就算存储在SSD（2TB）磁盘上，也需要400台机器存储，何况要考虑更复杂的网络操作、图片加载到内存进行计算的巨大开销等等。为此，阿里研发设计了一种更高阶的AMS深度学习训练架构，AMS比传统的Parameter Server(PS)架构更高阶，具体细节可参见论文Ge et al, “Image Matters: Visually modeling user behaviors using Advanced Model Server” 。

用户行为ID+图像的多模态 [阿里2017]

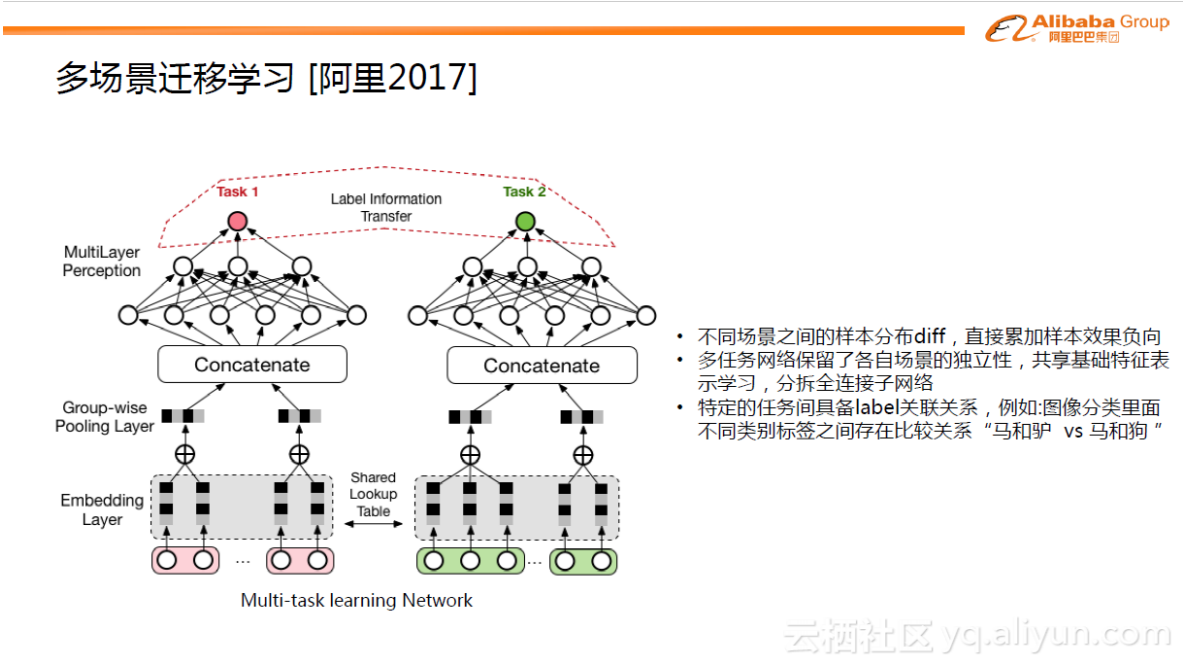


- 用户行为 和广告创意同时加入图片描述
- 训练涉及5亿图片，多达8T数据
- 100亿样本，500 ID/sample，平铺图片800TB
- 创新设计：带模型计算和更新的AMS架构
- 采用深层神经网络（16层）提取图片特征，部分图像网络参与训练，效率vs效果的tradeoff

Ge et al, “Image Matters: Visually modeling user behaviors using Advanced Model Server”

b.多场景迁移学习

对于模型而言，如果有更多的数据进行模型训练，效果一般都能得到提高。在手机淘宝端，我们有很多不同场景的广告位，如首页焦点图，导购场景等等。每个场景蕴含了用户的不同兴趣的表达。将不同场景直接进行合并用来训练模型，结果不是很乐观。因为不同场景之间的样本分布存在diff，直接累加样本会导致效果负向。

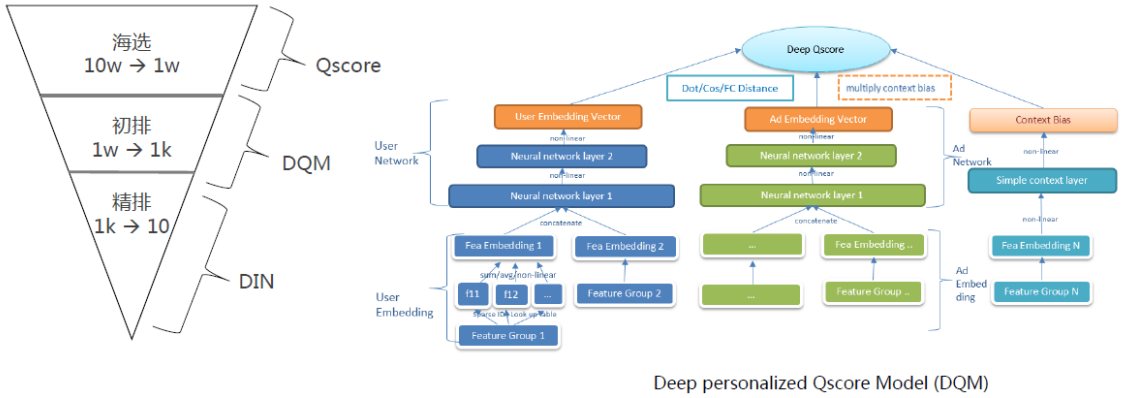


随着深度学习发展，发现用Multi-task learning(MTL)的方式可以很漂亮的解决这个问题。如上图中左边的例子，分为两个task，即分为两个子网络，对于底层的难以学习的Embedding层（或称为特征稀疏表达层）做了表示学习的共享（Shared Lookup Table），这种共享有助于大样本的子任务帮助小样本的子任务，使得底层的表达学习更加充分。对于上层的子网络，不同的task是分拆为不同的子网络，这样每个子网络可以各自去拟合自己task对应的概念分布。当然，在一些更复杂的问题中，需要进一步考虑不同task之间存在的关系，这种关系也可以帮助task达到更好的预测效果，这里叫做Label Information Transfer。MTL给跨场景迁移应用打开了一扇新的大门，可以充分的挖掘不同场景之间的独立性和关联性，从而帮助每个场景达到更好的效果。

c.深度个性化质量分网络

当用户在访问手机淘宝时，一瞬间，系统会有数千万的候选广告可以展现给用户，那具体展现哪些广告？下图中有简单的筛选过程。最开始通过一层匹配算法，圈出10万量级广告。这些广告需要在几十毫秒内展现给用户，如果全部进行复杂的模型打分，计算量是无法想象的，所以一般是分阶段进行：第一步利用简单的质量分数进行海选（Qscore是对每个Ad点击率的简单度量）。第二步利用DQM模型进行初排，这是一个从1万到1千的筛选过程。最后，用最复杂精细的模型，如DIN ，从1千中获取10个非常精准的广告。

深度个性化质量分网络 [阿里2017]



Deep personalized Qscore Model (DQM)

在第二步中，因为需要在几个毫秒内完成近万广告的打分过程，所以这个模型结构不能过于复杂。DQM模型类似与DSSM模型，分成不同域，如用户域，广告域以及场景域。将这些域表达成向量形式，最后的输出是通过向量间的简单操作，如内积操作，生成分数。相比传统的静态质量分Qscore模型，DQM引入了个性化，所以比Qscore好很多。

三、总结与展望——新的起点

1.怎么看深度学习技术？

我们认为，深度学习技术有三点优势。第一点，模型设计组件化。组件化是指在构建模型时，可以更多的关注idea和motivation本身，在真正数学化实现时可以像搭积木一样进行网络结构的设计和搭建。第二点，优化方法标准化。在2010年以前，Machine Learning还是一个要求较高的领域。它要求不仅了解问题、能定义出数学化的formulation，而且需要掌握很好的优化技巧，针对对应的问题设计具体的优化方法。但是在现在，深度学习因为模型结构上的变化，使得工业界可以用标准的SGD或SGD变种，很轻松的得到很好的优化解。第三点，深度学习可以帮助我们实现设计与优化的解耦，将设计和优化分阶段进行。对于工业界的同学来说，可以更加关注从问题本身出发，抽象和拟合领域知识。然后用一些标准的优化方法和框架来进行求解。

我们怎么看深度学习技术

优势一

模型设计组件化

优势二

优化方法标准化

优势三

设计与优化解耦

2.变革与展望

正是因为上面的优势，在最近两年内，整个工业界，包括整个阿里的广告体系里面，产生了革命性的技术模式变革。作为影响到广告营收的核心技术，CTR预估一直是互联网公司研究的焦点，很多公司都投入了大量的人力和物力进行研发。以前大家更多的是从特征工程的角度，结合人的先验去挖掘比较好的交叉组合特征。现在随着深度学习的引入，模式发生了巨大的转换。我提了一个新的概念，叫模型工程，现在我们可以用模型的方式以更少的人力、更高效的方式进行模式挖掘。这种变革在2015年到2017年间，在工业界的领先公司内，都成为了比较普遍的趋势。尤其在阿里内部，几乎是用比较野蛮的方式去革新了广告系统中方方面面的算法技术，不仅仅对CTR预估算法，还有匹配召回算法，以及一些机制和决策模型。现在，阿里认为深度学习处于V2.0时代。V1.0是属于掘金时代，大家都认为深度学习非常好用，野蛮式的发展。在V2.0时代，深度学习进入了精耕细作时代，当深度学习变成基础设施之后（比如微信已经成为日常生活中的工具），大家便可以利用这个工具，结合领域知识，更好的进行打磨并创新。